Distr.: General
27 March 2012

Original: English

Economic Commission for Europe

Conference of European Statisticians

UNECE-Eurostat Expert Group Meeting on Censuses Using Registers

Geneva, 22-23 May 2012
session 3: Availability, completeness and quality of data from registers and other sources

Register-Based Census 2011 in Slovenia – Some Quality Aspects

Note by Statistical Office of the Republic of Slovenia [1]

Summary

The first register-based Census in Slovenia was a success story from several points of view. In a time of fiscal restraint the budget savings around 14 million EUR are not negligible. The whole statistical process was divided into three stages following the availability of sources with first releases of data just 4 months (population), 6 months (households and families) and 12 months (other topics) after the reference date of 1 January 2011. The paper is dealing with some quality issues such as importance of cooperation and feedback between data keepers and Statistical Office or indirect measuring of over-registration. Despite two new data sources first time used in the statistical process (Household Register, Real Estate Register) the rate of missing data is in general lower than in the last field enumeration in 2002. Several quality indicators is also presented and explained. The register-based census is at the end figuratively a simple push on the button but after a very complex and demanding methodological and IT work with obligatory trial (in our case two years) and constant evaluation of all solutions.

[1] Prepared by Danilo Dolenc, Census Project Manager.

I. Background

1. Use of the registers and other administrative records in the statistical production has a long tradition in Slovenia. Following the long-term strategy to implement the Nordic model of statistics has resulted in the development of the legal bases for the register-based approach already in the early 1970s. As there were no similar initiatives in the other governmental bodies, the Statistical Office itself as a producer established four basic registers in close cooperation with corresponding authorities:

- Central Population Register (hereinafter the CPR) was established prior to the 1981 Census; since 1986 it has been used for production of population statistics;
- Register of Spatial Units (covering also address list) was set up in the 1980s;
- Statistical Register of Employment – from 1986 on it is updated monthly with data on pension, disability and health insurance provided by the Health Insurance Institute of Slovenia;
- Business Register has been available since 1976.

Three of the four registers (with the exception of the Statistical Register of Employment) are now kept by the administrative bodies.

2. Data from above-mentioned registers were used already at the 1991 Census and to a greater extent at the 2002 Census. The organization of the 2002 Census field enumeration and the statistical processing system were important steps to the complete register-based approach in the 2011 Census. Some contents were entirely taken over from the registers at the 2002 Census and therefore for the first time in censuses in Slovenia not included in the questionnaire and not collected in the field (e.g. place of birth, last migration, citizenship, marital status, activity status, occupation, industry, place of work). The imputation was the methodological solution for entries not existing in databases but found in the field enumeration.

3. From a content point of view the decision to go for a register-based census was possible because two missing sources became available after 2002:

- Real Estate Register (hereinafter the RER) was established in 2007 by the Surveying and Mapping Authority of Slovenia;
- Household Register with data on household composition as a part of the CPR was computerized (having for a long time been only in paper form and not used for statistical purposes).

Besides, the CPR addresses for residents in multi-dwellings buildings were supplemented with dwelling numbers from the RER.

II. Testing the data and feedback to data keepers

4. The next step after evaluation of methodological solutions and approval at the adequate body at the Office (Project Council) was the so-called ‘trial census’ with the primary goal to analyse and evaluate the quality of the input data, above all from the sources used for the first time in the statistical process. What proved to be more important during the census project is the mutual trust and good will of staff from all institutions involved. SORS reached the agreement that all partners put the priority to the quality improvement of sources needed for the census. As quality assessments and implementing quality assurance depend to a large extent on the quality of input data, the recognized inconsistencies have to be suppressed already in the administrative sources. SORS focused

most of its efforts on the preparation phase to persuade the data providers that the responsibility for the proper results is common.

5. Three key quality obstacles were recognized at this early stage connected with inconsistencies regarding household composition, under-coverage of dwelling numbers in the CPR and general unsatisfactory quality of housing data from the RER.

6. A comprehensive paper on quality assessment in the case of household data was presented at the European Conference on Quality in Official Statistics[2] in Helsinki in 2010. Most of the inconsistencies were very easy to solve by using matrixes on relation between parents-children and spouses in the household by using personal identification numbers (SID). SORS initiated methodological support to data providers and produced a very comprehensive document about lack of quality in the Household Register with detailed description of data errors and proposals of solutions.

7. The share of not useful data on relation to the reference person (unknown statement in the Household Register) was decreased from 31,000 in testing (1.5% of records) to fewer than 8,000. The errors found in combination of relation to the reference person and age (the most common mistakes were inverted relations to the reference person) has been almost entirely supressed (3,000 at testing).

8. The completeness of updating the dwelling number in the CPR was far below SORS's expectations as this variable is a key for matching dwellings with households. At the 'trial census' for about 400,000 persons (more than half of the population living in multi-dwelling buildings) those data were missing. This was a result of poor coordination between two register keepers and lack of awareness about the importance of those data for the statistical purposes.

9. Two main activities were undertaken in close cooperation between the Ministry of the Interior and the Statistical Office:

- Methodological solutions for automated determination of missing dwelling numbers by linkage of data on ownership of dwellings, registered residence of owners and their households. Dwelling numbers were automatically allocated to the CPR records in the case of complete matching (the presumption was that most of the owners live in their properties);
- After finishing the previous action an official letter was sent to the reference person of the household living in multi-dwelling buildings but still without a dwelling number to report (free of charge) the needed data (with financial, organizational and processing support of the Statistical Office). 49,000 letters were sent and the total response rate was 75% (including returning letters because the person is unknown at the address or has already moved to another address). From the quality point of view the inquiry was of special importance for the administrative body as respondents sent also a lot of comments.

10. In case of the RER we got in advance preliminary data and SORS's statistical analyses helped administrative staff to focus work on quality improvement. The RER is a completely new source based on field data collection (a kind of a housing census). The most disputable housing variables were ownership (long-lasted administrative procedures), the number of rooms in connection to useful floor space, and seasonal or secondary use.

[2] http://q2010.stat.fi/sessions/session-26/

III. Coherence with regular population statistics and measuring over-registration

11. In order to maintain the coherence with other population statistics, for the first time the census definition is harmonized with regular (quarterly) population statistics. In fact the pre-defined and statistically processed population data from only one administrative source (CPR) prepared according to the EU Regulation on migration from 2007 are the basic input database in the register-based census. Because of the improved regular statistical process, the extraction from the source started at the beginning of April 2011 (three months delay from the reference date of 1 January) and therefore the quality of administrative data is higher and data are more accurate. The first stage of the process (integration of input data for population, households and housing) resulted in the first release of final census data on basic population topics at the end of April 2011.

12. Register data are as a rule administrative records which do not necessarily correspond to statistical concepts but depend on legislation. Besides, formally correct data in the register do not necessarily correspond to the real situation of population and households. The quality from this point of view depends on how strictly people respect the legislation on registration.

Administrative concepts	Statistical concepts
More than one registered residence in Slovenia at the same time possible: one permanent and/or one or more temporary residence Temporary absence abroad (but still with registered residence in Slovenia)	Only one (usual) residence at the same time per person: ♦ length and intention of stay (one-year criterion) ♦ priority rules in the case of two administratively registered residences

13. Over-registration in the registers is a common problem of all register-based systems. Direct measuring is impossible. On the basis of other field surveys we estimated over-registration up to 1%. The Register-based Census 2011 gives for the first time a very precise answer about the number of persons defined as population according to the statistical methodology but already moved out of the country. We concentrate our analyses on persons with permanent residence in Slovenia only as we have data on expiry date of the permit for temporary residing in Slovenia for foreigners and we assume that the person emigrated if there is no prolongation of the permit.

14. As an indicator of still living in Slovenia at the reference date we use data on activity status. Eight different sources were used for producing data on activity status by using the source hierarchy methodology[3]. The sources content are employed and unemployed persons, full-time and part-time students, scholarship recipients, pension recipients, recipients of social benefits, income tax payers and all persons in health insurance including family members of insured persons.

15. The main findings are:

- for 1.25% of the population no evidence in any activity status data source;

[3] http://www.stat.si/popis2011/eng/MP_Akt.aspx?lang=eng

- but for 37% of those records data on educational attainment exist at least in one of the sources[4] but those data refer to a longer period;
- the highest levels of over-registration are recorded for foreigners with permanent residence in Slovenia (6.1%), for working age population aged 30-44 years (2.0%) and for population over 94 years (4.7% - so called administrative survivors).

16. If summarized, the results are more or less as expected and taking into account other factors that might have effect on proper linkage of data the real over-registration is estimated at 0.9%. Due to the comparable results with other statistical household surveys there is no need for additional post-enumeration survey focused on over-registration.

IV. Linkage of data on persons, households and dwellings

17. Linkage of data on persons, households and dwellings using unique identifiers is one of the most important tasks producing census data by field enumeration or by using register data. In case of the Slovenian Register-based Census 2011 the direct linkage of all data sources for persons using PIN was the basic statistical operation as the PIN is the key in all databases. The PIN in all basic databases has been transformed before the beginning of the statistical process to the indefinable statistical identifier (SID).

18. The household identifier in the Household Register is the serial number of the household running from 1 to NNN at the same address. The most important advantage of the Household Register compared to other countries that use registers is the availability of data on relation to the reference person of the household. Recommendations (2006) distinguish two concepts of private households (paragraphs 479-481): the household-dwelling concept (exactly one household per occupied housing unit) and the housekeeping concept (two or more households can share the same dwelling). In most countries that use registers the first concept is applied but in Slovenia the existing Household Register enables us to implement the housekeeping concept, which is in fact more appropriate as cohabitation of more households in the same dwelling (mostly in detached houses) is significant in Slovenia.

19 The main quality obstacle in terms of household identifier is the fact that household ID (and also relation to the reference person of the household) is available only for permanent residence and consequently data are not available for more than half of the foreigners and also for Slovenian citizens having statistically usual residence at the temporary address. After integration of household and person data, 2.1% of household ID's were missing (94% for foreigners).

20. Besides, 36,000 records on relation to the reference person of the household, which was the key variable for statistical family's generation, were not directly useful for these purposes. Relation to the reference person of the household was predicted to be the third important key identifier.

21. Despite all action undertaken for improving quality on dwelling number for persons living in multi-dwelling buildings, the share of missing dwellings ID's was still high (89,000 or 12.3% of considered population). Besides, comparing RER and CPR data we found out that additional 16,000 ID's in the CPR do not correspond to the RER ID's.

[4] http://www.stat.si/popis2011/eng/MP_Izo.aspx?lang=eng

22. Very detailed methodological rules were prepared followed by complex IT solutions for imputing missing key identifiers or their corrections. Having the quality of household and family data in mind, we decided to build in the statistical process an interface for manual editing after a very complex automated process for generating ID's. Two main problems had to be solved by manual editing: connecting children to their parents in case of foreigner's households and family formation in households with several members where automation is not rational (manual coding of family status). The interface which also includes the surnames allows correction of very few strictly approved variables.

23. The manual editing was done in 7 full-time working days by 5 employees of the Statistical Office. Even though only 20,000 (1%) records were edited manually with an average of two corrections per record, the quality impact was of great importance for us having in mind dissemination of data to very detailed geographical levels.

24. The distribution of input and output data for key identifiers in the whole statistical process is presented in Table 1.

Table 1 Quality indicators for key identifiers

Identifier	Number of records	Unchanged	Imputed		Corrections	
			Automated	Manual	Automated	Manual
		Share in %				
Dwelling ID[1)]	724,479	75.3	11.7	0.6	11.9	0.5
Household ID[2)]	2,016,423	94.9	2.0	0.1	2.3	0.6
Relation to the reference person[2)]	2,016,423	91.6	4.1	0.1	3.3	0.8[3)]

[1)] Multi-dwelling buildings only. [2)] Private households only. [3)] Manual corrections in the stage of family generation also included.

V. Family generation

25. The family concept used in the Register-based Census 2011 is harmonized with Recommendations (2006, paragraph 493) and with EU Regulation on population and housing censuses. Family generation is pure statistical operation in principle based on data on:

- Relation between biological parents and children and between married couples in the household using SID's of mother, father and spouse;
- Relation to the reference person of the household in case of missing links from the previous point.

26. The quality of family data is directly connected to the quality of household data. The main obstacles for family generation were:

- SID's of parents and spouses are not available for older generations and for most foreigners;
- Some relations to the reference person are not identifying enough (other relative, non-relative);

- Other relative often means the cohabiting partner of one of the household members (except the reference person);
- In case of more than one family in the household, household members could belong to every family;
- Complexity of relations in households with many members.

27. Four main methods were used for family generation:

1) Automated family generation by using SID's with automated correction of eventual errors of relations to the reference person

Code	Relation to the reference person	SID	SID of parent	Family
0	Reference person	1	8	1
3	Child	2	1	2
3	Child	3	1	1
9	Grandchild	4	2	2

2) Automated family generation using relation to the reference person matrix. 50 basic unique household matrixes have been developed with 212 sub-matrixes. Sub-matrixes are needed because of different combinations of relations in the household within the same type of family. For the most common and simple type of household (one-family household husband-wife couple with children without other persons) three sub-matrixes had to be produced (see below).

Sub-matrix 1		Sub-matrix 2		Sub-matrix 3	
Code	**Relation to the reference**	**Code**	**Relation to the reference**	**Code**	**Relation to the reference**
0	Reference person	0	Reference	0	Reference person
1	Spouse	5	Parents (both of	5	Parents (both of
3	Child*			10	Sibling*

*Number of children in sub-matrix 1 and number of siblings in sub-matrix 3 are not limited.

3) Manual family generation using interface in case of partly automated generation of families because of some missing SID's and with not identifying relations to the reference persons.

4) Manual corrections of family data in case of using matrixes as some relations were not coded properly in the administrative source (e.g. child coded as parent or vice-versa).

28. Of the 547,000 multi-person private households with possible existent family only for less than 0.5% of households the manual family generation applied. At the last stage of processing of family data approximately 50 rules were set up for checking inconsistencies in age or family status between family members and manual editing was presumed a better solution in terms of quality and time used for this part of the statistical process instead of automated corrections due to several possible errors with very unique solutions. Errors were found in 3,700 households (0.7%) with 14,000 members, but only 6,000 records were corrected.

29. Family generation was one of the most demanding, time consuming and questionable statistical processes in terms of quality in the previous field censuses but in the register-based census this operation was executed almost perfectly.

VI. Economic and educational variables – complexity of sources

30. The most different sources were used for production of economic and educational data in the Register-based Census 2011 (8 sources for activity and 9 sources for educational attainment). We examined each provided source of the data very accurately and assessed suitability for use in terms of quality, reliability, timeliness, accessibility and comparability. The basic methodological principle is the hierarchy of the sources, which means that in case of availability of data in several sources, the priority is given to the source indicated with higher priority (irrespective of whether data in different sources are the same or different).

31. Preparation of the data on educational attainment is a complex statistical process due to the collection of data on persons who participated in education in different periods and different educational systems. It should also be noted that an increasing part of the population is still in the educational process at the time of data collection. On the other hand, as a rule, data on activity refer to 1 January 2011, where we have assumed at some individual source that data refer to this time point.

32. The main sources of data on activity status were the Statistical Register of Employment (45%) and Health Insurance Institute of Slovenia data on family members of insured persons and other inactive persons in health insurance (29%). The main sources of data on educational attainment were again the Statistical Register of Employment (56%) and Census 2002 data mostly for oldest population (13%). Despite so many sources used, there are still missing data which had to be imputed. All missing data except data on occupation and industry for population working abroad because there is no donor available are imputed. The main statistical method used for imputation of missing values is the hot deck imputation method. Imputation rates for obligatory topics according to the Regulation on population and housing censuses are:

- Activity status 1.5%
- Occupation (employed) 3.9%
- Occupation (unemployed) 5.2%
- Industry (employed) 3.7%
- Industry (unemployed) 18.0%
- Status in employment (employed) 3.7%
- Place of work 3.8%
- Educational attainment 1.5%

33. Imputation rates for comparable variables are lower than imputation rates in our last field enumeration in 2002 (e.g. educational attainment 2.4%). Besides, the whole statistical process is under controlled and uniform methodological approach in all stages of the process compared to the thousands of enumerators and several dozen data editors in the previous statistical process.

VII. Conclusion

34. The Register-based Census 2011 in Slovenia was a success story carried out by own human resources without any outsourcing. A small project team composed of six employees was responsible for the whole methodological work and statistical process. Continuity in relation to cooperation with register keepers at all levels (managerial, statistical experts, technical) and proactive role of the Statistical Office are crucial elements to preserve the possibility for future register-based censuses. Besides, the permanent and long-continued use of the administrative data in the statistical process is one of the most important prerequisites for better quality. From this point of view the new sources used for the first time in the statistical process (such as the Household Register and the Real Estate Register) will be definitely improved in the future as the feedback is given to the register keepers who are aware of their responsibility towards data users at different levels. We statisticians have a privilege to modify administrative data according to the statistical concepts and we can also use statistical methods but for the administrative use this is not allowed without cooperation of the register subjects.

35. We are planning to produce complete census results every three to four years (two times in between regular decennial censuses). Some topics besides basic demographic variables (which are produced quarterly) – such as educational attainment and activity status – will be from 2011 onwards available for our users annually using the same methodology as in the Register-based Census 2011.

VIII. References

Dolenc D. (2010), "Quality Assessment in Register-based Census – Administrative versus Statistical Concepts in the Case of Households", European Conference on Quality in Official Statistics, Helsinki, Finland

Dolenc D. (2010), "First register-based census in Slovenia: How to convert administrative sources to statistics", Statistics Canada Symposium 2010, Ottawa, Canada

Recommendations for the 2010 Censuses of Population and Housing, United Nations, New York and Geneva, 2006

Register-based Census 2011 website, Statistical Office of the Republic of Slovenia,

http://www.stat.si/popis2011/eng/Default.aspx?lang=eng